

The multilocus sequence typing network: mlst.net

David M. Aanensen* and Brian G. Spratt

Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Hospital, London W2 1PG, UK

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

The unambiguous characterization of strains of a pathogen is crucial for addressing questions relating to its epidemiology, population and evolutionary biology. Multilocus sequence typing (MLST), which defines strains from the sequences at seven house-keeping loci, has become the method of choice for molecular typing of many bacterial and fungal pathogens (and non-pathogens), and MLST schemes and strain databases are available for a growing number of prokaryotic and eukaryotic organisms. Sequence data are ideal for strain characterization as they are unambiguous, meaning strains can readily be compared between laboratories via the Internet. Laboratories undertaking MLST can quickly progress from sequencing the seven gene fragments to characterizing their strains and relating them to those submitted by others and to the population as a whole. We provide the gateway to a number of MLST schemes, each of which contain a set of tools for the initial characterization of strains, and methods for relating query strains to other strains of the species, including clustering based on differences in allelic profiles, phylogenetic trees based on concatenated sequences, and a recently developed method (eBURST) for identifying clonal complexes within a species and displaying the overall structure of the population. This network of MLST websites is available at <http://www.mlst.net>

INTRODUCTION

Multilocus sequence typing (MLST) is a nucleotide sequence-based approach to the unambiguous characterization of strains of bacterial species, or other microbial species, via the Internet (1,2). MLST involves obtaining the sequences of internal fragments of seven house-keeping genes for each strain of a particular species. The sequences of each fragment are compared with all the previously identified sequences (alleles) at that locus and, thereby, are assigned allele numbers at each of the

seven loci. The combination of the seven allele numbers defines the allelic profile of the strain and each different allelic profile is assigned as a sequence type (ST), which is used to describe the strain.

Nucleotide sequencing is relatively cheap, and easy to perform. The data produced by MLST are ideal for the characterization of strains of bacterial or fungal species via a web server. MLST is now widely used for molecular epidemiology as it allows strains studied by different groups to be compared and MLST schemes have been developed for ~20 bacteria (mostly pathogens) (3), and three fungi (4,5) and databases that can be queried have been available for several years (6). The MLST databases are currently hosted on two main web servers located at Imperial College London (<http://www.mlst.net>) and Oxford University [<http://pubmlst.org>; (7)]. The former web server acts as a gateway to a number of species-specific websites each of which contains tools for the analysis of allele sequences and STs, and a web interface for obtaining epidemiological information held on the increasing numbers of strains that are submitted by the user community.

Along with centrally available tools for those interested in starting their own MLST schemes, such as for defining alleles using non-redundant databases (NRDB), measuring linkage disequilibrium and an interface to Splits Tree (8), <http://www.mlst.net> provides a number of options to display the relatedness of query strains to those in the strain database.

MLST WORKFLOW

Laboratories undertaking MLST can access species-specific information on each of the individual mlst.net species websites, including sequencing protocols and primer sequences, allowing a laboratory to begin producing data rapidly. Characterization of a strain requires the generation of the sequences of the seven gene fragments and, once these are available, they are used to query the appropriate mlst.net website, to assign the alleles at each locus and thereby to obtain the allelic profile of the query strain. Each MLST website holds the sequences of all known alleles at each of the seven loci, and all known allelic profiles, and through the curator assigns new allele numbers and STs. Every different sequence at each locus is assigned as a distinct allele and new alleles are assigned allele numbers by the curator and are entered in the allele database.

*To whom correspondence should be addressed. Tel: +44 0 20 7594 3825; Fax: +44 0 20 7594 3693; Email: d.aanensen@imperial.ac.uk

A Single Locus Query

Please choose the allele you wish to query -

gdh_

Multiple Locus Query

Please enter sequences below

areo

gdh_

Simple Results

CTTCTTTGAGACGATGCGATTCGCTGGTGGTCTCTTTTCGGTACAGGTAACG
 ACTGACTGAAAGAGGACCCATGTCACACATGCTTTTAAACAAATGGATTCATCTTGG
 AGAACCACTTGTCCCAATTTTACCACTATTTTAAACAAATGGATTCATCTTGG
 TAGCTCAATGGGAGAGAGATTTTAAAGTGGTCTCTTTTCGGTACAGGTAACG
 CGGTACAGATGCGACTG
 ATACAGATGCGAGGAGATTTTAAAGTGGTCTCTTTTCGGTACAGGTAACG
 ATACAGATGCGAGGAGATTTTAAAGTGGTCTCTTTTCGGTACAGGTAACG
 GGGTCACTGCAATTTTGGTCTCTTTTCGGTACAGGTAACG
 CTGACAGATTTTGGTCTCTTTTCGGTACAGGTAACG

gdh_

Submit

B Single Locus Query - Results

SEARCH RESULT: Sequence not found, closest similarity is 99% allele - 1
 You may have a new allele.

Locus Allele Number	Error Messages	Action
areo_1	OK	None
gdh_1	Not found	Check Base differences
gdh_2	OK	None
gdh_3	Closest similarity is 99%, may be new allele	None
recp_1	OK	None
spl_1	OK	None
spl_2	OK	None
spl_3	OK	None
spl_4	OK	None
spl_5	OK	None
spl_6	OK	None
spl_7	OK	None
spl_8	OK	None
spl_9	OK	None
spl_10	OK	None
spl_11	OK	None
spl_12	OK	None
spl_13	OK	None
spl_14	OK	None </tr

Submit data for allelic profile query

areo

gdh_

recp

spl_

ddl_

Exact or nearest match

Submit

C Multiple Locus Query - Results

SEARCH RESULT: Sequence not found, closest similarity is 99% allele - 1
 You may have a new allele.

Following is a list of nucleotide differences between your sequence and other gdh_ alleles in the database. Click here for help and explanation

position 9
 query AACACTTATCCG
 gdh_1 AACACTTATCCG

1 C T A T C d=1 s=99.8%
 2 C T A T C d=2 s=99.6%
 3 C T A T C d=2 s=99.8%
 4 C T A T C d=2 s=99.8%
 5 C T A T C d=2 s=99.8%
 6 C T A T C d=2 s=99.8%
 7 C T A T C d=2 s=99.8%
 8 C T A T C d=2 s=99.8%
 9 C T A T C d=2 s=99.8%
 10 C T A T C d=2 s=99.8%
 11 C T A T C d=2 s=99.8%
 12 C T A T C d=2 s=99.8%
 13 C T A T C d=2 s=99.8%
 14 C T A T C d=2 s=99.8%
 15 C T A T C d=2 s=99.8%
 16 C T A T C d=2 s=99.8%
 17 C T A T C d=2 s=99.8%
 18 C T A T C d=2 s=99.8%
 19 C T A T C d=2 s=99.8%
 20 C T A T C d=2 s=99.8%
 21 C T A T C d=2 s=99.8%
 22 C T A T C d=2 s=99.8%
 23 C T A T C d=2 s=99.8%
 24 C T A T C d=2 s=99.8%
 25 C T A T C d=2 s=99.8%
 26 C T A T C d=2 s=99.8%
 27 C T A T C d=2 s=99.8%
 28 C T A T C d=2 s=99.8%
 29 C T A T C d=2 s=99.8%
 30 C T A T C d=2 s=99.8%
 31 C T A T C d=2 s=99.8%
 32 C T A T C d=2 s=99.8%
 33 C T A T C d=2 s=99.8%
 34 C T A T C d=2 s=99.8%
 35 C T A T C d=2 s=99.8%
 36 C T A T C d=2 s=99.8%
 37 C T A T C d=2 s=99.8%
 38 C T A T C d=2 s=99.8%
 39 C T A T C d=2 s=99.8%
 40 C T A T C d=2 s=99.8%
 41 C T A T C d=2 s=99.8%
 42 C T A T C d=2 s=99.8%
 43 C T A T C d=2 s=99.8%
 44 C T A T C d=2 s=99.8%
 45 C T A T C d=2 s=99.8%
 46 C T A T C d=2 s=99.8%
 47 C T A T C d=2 s=99.8%
 48 C T A T C d=2 s=99.8%
 49 C T A T C d=2 s=99.8%
 50 C T A T C d=2 s=99.8%
 51 C T A T C d=2 s=99.8%
 52 C T A T C d=2 s=99.8%
 53 C T A T C d=2 s=99.8%
 54 C T A T C d=2 s=99.8%
 55 C T A T C d=2 s=99.8%
 56 C T A T C d=2 s=99.8%
 57 C T A T C d=2 s=99.8%
 58 C T A T C d=2 s=99.8%
 59 C T A T C d=2 s=99.8%
 60 C T A T C d=2 s=99.8%
 61 C T A T C d=2 s=99.8%

D Batch Strain Query

Choose an XML file to upload - see help for XML format.

C:\strain_sequences.xml

Submit

E Batch Strain Query - Results

Strain	areo	gdh	recp	spl	ddl	ST
Strain1	1	1	1	1	1	incomplete
Strain2	1	1	1	1	1	incomplete
Strain3	1	1	1	1	1	incomplete
Strain4	1	1	1	1	1	incomplete
Strain5	1	1	1	1	1	incomplete
Strain6	1	1	1	1	1	incomplete
Strain7	1	1	1	1	1	incomplete
Strain8	1	1	1	1	1	incomplete
Strain9	1	1	1	1	1	incomplete
Strain10	1	1	1	1	1	incomplete
Strain11	1	1	1	1	1	incomplete
Strain12	1	1	1	1	1	incomplete
Strain13	1	1	1	1	1	incomplete
Strain14	1	1	1	1	1	incomplete
Strain15	1	1	1	1	1	incomplete

View unknown STs with eBURST

F Allelic Profile Query

Please enter your query below (the figure below each box represents the number of unique alleles in the current database for that locus).

areo

gdh_

recp

spl_

ddl_

Submit

G Allelic Profile Query - Results

Please select Query type:

4 or more matches

70 101 113 125 137 149 161 173 185 197 209 221 233 245 257 269 281 293 305 317 329 341 353 365 377 389 401 413 425 437 449 461 473 485 497 509 521 533 545 557 569 581 593 605 617 629 641 653 665 677 689 701 713 725 737 749 761 773 785 797 809 821 833 845 857 869 881 893 905 917 929 941 953 965 977 989 1001 1013 1025 1037 1049 1061 1073 1085 1097 1109 1121 1133 1145 1157 1169 1181 1193 1205 1217 1229 1241 1253 1265 1277 1289 1301 1313 1325 1337 1349 1361 1373 1385 1397 1409 1421 1433 1445 1457 1469 1481 1493 1505 1517 1529 1541 1553 1565 1577 1589 1601 1613 1625 1637 1649 1661 1673 1685 1697 1709 1721 1733 1745 1757 1769 1781 1793 1805 1817 1829 1841 1853 1865 1877 1889 1901 1913 1925 1937 1949 1961 1973 1985 1997 2009 2021 2033 2045 2057 2069 2081 2093 2105 2117 2129 2141 2153 2165 2177 2189 2201 2213 2225 2237 2249 2261 2273 2285 2297 2309 2321 2333 2345 2357 2369 2381 2393 2405 2417 2429 2441 2453 2465 2477 2489 2501 2513 2525 2537 2549 2561 2573 2585 2597 2609 2621 2633 2645 2657 2669 2681 2693 2705 2717 2729 2741 2753 2765 2777 2789 2801 2813 2825 2837 2849 2861 2873 2885 2897 2909 2921 2933 2945 2957 2969 2981 2993 3005 3017 3029 3041 3053 3065 3077 3089 3101 3113 3125 3137 3149 3161 3173 3185 3197 3209 3221 3233 3245 3257 3269 3281 3293 3305 3317 3329 3341 3353 3365 3377 3389 3401 3413 3425 3437 3449 3461 3473 3485 3497 3509 3521 3533 3545 3557 3569 3581 3593 3605 3617 3629 3641 3653 3665 3677 3689 3701 3713 3725 3737 3749 3761 3773 3785 3797 3809 3821 3833 3845 3857 3869 3881 3893 3905 3917 3929 3941 3953 3965 3977 3989 4001 4013 4025 4037 4049 4061 4073 4085 4097 4109 4121 4133 4145 4157 4169 4181 4193 4205 4217 4229 4241 4253 4265 4277 4289 4301 4313 4325 4337 4349 4361 4373 4385 4397 4409 4421 4433 4445 4457 4469 4481 4493 4505 4517 4529 4541 4553 4565 4577 4589 4601 4613 4625 4637 4649 4661 4673 4685 4697 4709 4721 4733 4745 4757 4769 4781 4793 4805 4817 4829 4841 4853 4865 4877 4889 4901 4913 4925 4937 4949 4961 4973 4985 4997 5009 5021 5033 5045 5057 5069 5081 5093 5105 5117 5129 5141 5153 5165 5177 5189 5201 5213 5225 5237 5249 5261 5273 5285 5297 5309 5321 5333 5345 5357 5369 5381 5393 5405 5417 5429 5441 5453 5465 5477 5489 5501 5513 5525 5537 5549 5561 5573 5585 5597 5609 5621 5633 5645 5657 5669 5681 5693 5705 5717 5729 5741 5753 5765 5777 5789 5801 5813 5825 5837 5849 5861 5873 5885 5897 5909 5921 5933 5945 5957 5969 5981 5993 6005 6017 6029 6041 6053 6065 6077 6089 6101 6113 6125 6137 6149 6161 6173 6185 6197 6209 6221 6233 6245 6257 6269 6281 6293 6305 6317 6329 6341 6353 6365 6377 6389 6401 6413 6425 6437 6449 6461 6473 6485 6497 6509 6521 6533 6545 6557 6569 6581 6593 6605 6617 6629 6641 6653 6665 6677 6689 6701 6713 6725 6737 6749 6761 6773 6785 6797 6809 6821 6833 6845 6857 6869 6881 6893 6905 6917 6929 6941 6953 6965 6977 6989 7001 7013 7025 7037 7049 7061 7073 7085 7097 7109 7121 7133 7145 7157 7169 7181 7193 7205 7217 7229 7241 7253 7265 7277 7289 7301 7313 7325 7337 7349 7361 7373 7385 7397 7409 7421 7433 7445 7457 7469 7481 7493 7505 7517 7529 7541 7553 7565 7577 7589 7601 7613 7625 7637 7649 7661 7673 7685 7697 7709 7721 7733 7745 7757 7769 7781 7793 7805 7817 7829 7841 7853 7865 7877 7889 7901 7913 7925 7937 7949 7961 7973 7985 7997 8009 8021 8033 8045 8057 8069 8081 8093 8105 8117 8129 8141 8153 8165 8177 8189 8201 8213 8225 8237 8249 8261 8273 8285 8297 8309 8321 8333 8345 8357 8369 8381 8393 8405 8417 8429 8441 8453 8465 8477 8489 8501 8513 8525 8537 8549 8561 8573 8585 8597 8609 8621 8633 8645 8657 8669 8681 8693 8705 8717 8729 8741 8753 8765 8777 8789 8801 8813 8825 8837 8849 8861 8873 8885 8897 8909 8921 8933 8945 8957 8969 8981 8993 9005 9017 9029 9041 9053 9065 9077 9089 9101 9113 9125 9137 9149 9161 9173 9185 9197 9209 9221 9233 9245 9257 9269 9281 9293 9305 9317 9329 9341 9353 9365 9377 9389 9401 9413 9425 9437 9449 9461 9473 9485 9497 9509 9521 9533 9545 9557 9569 9581 9593 9605 9617 9629 9641 9653 9665 9677 9689 9701 9713 9725 9737 9749 9761 9773 9785 9797 9809 9821 9833 9845 9857 9869 9881 9893 9905 9917 9929 9941 9953 9965 9977 9989 10009 10021 10033 10045 10057 10069 10081 10093 10105 10117 10129 10141 10153 10165 10177 10189 10201 10213 10225 10237 10249 10261 10273 10285 10297 10309 10321 10333 10345 10357 10369 10381 10393 10405 10417 10429 10441 10453 10465 10477 10489 10501 10513 10525 10537 10549 10561 10573 10585 10597 10609 10621 10633 10645 10657 10669 10681 10693 10705 10717 10729 10741 10753 10765 10777 10789 10801 10813 10825 10837 10849 10861 10873 10885 10897 10909 10921 10933 10945 10957 10969 10981 10993 11005 11017 11029 11041 11053 11065 11077 11089 11101 11113 11125 11137 11149 11161 11173 11185 11197 11209 11221 11233 11245 11257 11269 11281 11293 11305 11317 11329 11341 11353 11365 11377 11389 11401 11413 11425 11437 11449 11461 11473 11485 11497 11509 11521 11533 11545 11557 11569 11581 11593 11605 11617 11629 11641 11653 11665 11677 11689 11701 11713 11725 11737 11749 11761 11773 11785 11797 11809 11821 11833 11845 11857 11869 11881 11893 11905 11917 11929 11941 11953 11965 11977 11989 12001 12013 12025 12037 12049 12061 12073 12085 12097 12109 12121 12133 12145 12157 12169 12181 12193 12205 12217 12229 12241 12253 12265 12277 12289 12301 12313 12325 12337 12349 12361 12373 12385 12397 12409 12421 12433 12445 12457 12469 12481 12493 12505 12517 12529 12541 12553 12565 12577 12589 12601 12613 12625 12637 12649 12661 12673 12685 12697 12709 12721 12733 12745 12757 12769 12781 12793 12805 12817 12829 12841 12853 12865 12877 12889 12901 12913 12925 12937 12949 12961 12973 12985 12997 13009 13021 13033 13045 13057 13069 13081 13093 13105 13117 13129 13141 13153 13165 13177 13189 13201 13213 13225 13237 13249 13261 13273 13285 13297 13309 13321 13333 13345 13357 13369 13381 13393 13405 13417 13429 13441 13453 13465 13477 13489 13501 13513 13525 13537 13549 13561 13573 13585 13597 13609 13621 13633 13645 13657 13669 13681 13693 13705 13717 13729 13741 13753 13765 13777 13789 13801 13813 13825 13837 13849 13861 13873 13885 13897 13909 13921 13933 13945 13957 13969 13981 13993 14005 14017 14029 14041 14053 14065 14077 14089 14101 14113 14125 14137 14149 14161 14173 14185 14197 14209 14221 14233 14245 14257 14269 14281 14293 14305 14317 14329 14341 14353 14365 14377 14389 14401 14413 14425 14437 14449 14461 14473 14485 14497 14509 14521 14533 14545 14557 14569 14581 14593 14605 14617 14629 14641 14653 14665 14677 14689 14701 14713 14725 14737 14749 14761 14773 14785 14797 14809 14821 14833 14845 14857 14869 14881 14893 14905 14917 14929 14941 14953 14965 14977 14989 15001 15013 15025 15037 15049 15061 15073 15085 15097 15109 15121 15133 15145 15157 15169 15181 15193 15205 15217 15229 15241 15253 15265 15277 15289 15301 15313 15325 15337 15349 15361 15373 15385 15397 15409 15421 15433 15445 15457 15469 15481 15493 15505 15517 15529 15541 15553 15565 15577 15589 15601 15613 15625 15637 15649 15661 15673 15685 15697 15709 15721 15733 15745 15757 15769 15781 15793 15805 15817 15829 15841 15853 15865 15877 15889 15901 15913 15925 15937 15949 15961 15973 15985 15997 16009 16021 16033 16045 16057 16069 16081 16093 16105 16117 16129 16141 16153 16165 16177 16189 16201 16213 16225 16237 16249 16261 16273 16285 16297 16309 16321 16333 16345 16357 16369 16381 16393 16405 16417 16429 16441 16453 16465 16477 16489 16501 16513 16525 16537 16549 16561 16573 16585 16597 16609 16621 16633 16645 16657 16669 16681 16693 16705 16717 16729 16741 16753 16765 16777 16789 16801 16813 16825 16837 16849 16861 16873 16885 16897 16909 16921 16933 16945 16957 16969 16981 16993 17005 17017 17029 17041 17053 17065 17077 17089 17101 17113 17125 17137 17149 17161 17173 17185 17197 17209 17221 17233 17245 17257 17269 17281 17293 17305 17317 17329 17341 17353 17365 17377 17389 17401 17413 17425 17437 17449 17461 17473 17485 17497 17509 17521 17533 17545 17557 17569 17581 17593 17605 17617 17629 17641 17653 17665 17677 17689 17701 17713 17725 17737 17749 17761 17773 17785 17797 17809 17821 17833 17845 17857 17869 17881 17893 17905 17917 17929 17941 17953 17965 17977 17989 18001 18013 18025 18037 18049 18061 18073 18085 18097 18109 18121 18133 18145 18157 18169 18181 18193 18205 18217 18229 18241 18253 18265

Each MLST species website offers a number of analysis steps for a user. First, alleles have to be assigned from the sequence data by one of three options (Figure 1A and D):

Single/batch locus query: allowing a single sequence or a batch of sequences for a single locus to be compared with all known alleles.

Multiple locus query: allowing the input of the sequences of all seven loci for a single strain.

Batch strain query: allowing input of the sequences of all seven loci for a batch of strains.

In all cases, the user's sequence is checked for correct length for that locus, and for the absence of unexpected characters, and is then queried against all other sequences in the species database. For *Candida albicans*, a diploid organism, the standard ambiguity codes are allowed and are used to assign heterozygous nucleotide sites (4).

If the user's sequence is found, the allele number is returned, whereas if the user has a novel sequence, the percentage identity to the closest allele in the database is returned and the user is advised to check carefully those nucleotide sites that differ from the most similar allele or alleles in the database (Figure 1B). This can be carried out using the Jalview alignment editor (9), or the nucleotide differences can be displayed between the query sequence and the most similar alleles, as in Figure 1C. The latter method allows the user to view the flanking sequence around each nucleotide difference between the query and the most similar alleles, allowing the trace files of their proposed new allele to be searched easily for any potential ambiguities or sequencing errors. If a user is confident that they have a new allele, the forward and reverse trace files are submitted to the MLST scheme curator, as a quality control check, before a new allele number is assigned by the curator for the novel sequence. The sequence of the new allele is then entered into the database.

Repeating this process for each locus provides the seven-digit allelic profile for the query strain. The seven-digit allelic profile can then be entered into the allelic profile query to discover if the strain is identical, or similar, in allelic profile to any of the strains already in the database (Figure 1F). The multiple locus query represents a batch processing method for a single strain, allowing all seven sequences for the query strain to be entered at once and for the allele numbers at each locus to be returned.

BATCH STRAIN QUERY

The repeated querying of single strains becomes very time consuming for laboratories undertaking MLST on many hundred strains of a particular species. The need to analyze sequences from multiple genes from a large number of strains at the same time precludes the use of standard sequence

```
<dataroot>
<strain_sequences>
<strainid>STRAIN1</strainid>
<locus1>...GCCAGTGA...</locus1>
<locus2>...CCGAGTGA...</locus2>
<locus3>...GCGAGTGA...</locus3>
<locus4>...ACGGGTGA...</locus4>
<locus5>...TCGAGGGA...</locus5>
<locus6>...GCGAGTGA...</locus6>
<locus7>...CCGTGAGA...</locus7>
</strain_sequences>
<strain_sequences>
<strainid>STRAIN2</strainid>
<locus1>...GCCAGTGA...</locus1>
<locus2>...CCGAGTGA...</locus2>
<locus3>...GCGAGTGA...</locus3>
<locus4>...ACGGGTGA...</locus4>
<locus5>...TCGAGGGA...</locus5>
<locus6>...GCGAGTGA...</locus6>
<locus7>...CCGTGAGA...</locus7>
</strain_sequences>
etc.....
</dataroot>
```

Figure 2. The XML format for batch querying multiple strains using mlst.net.

formats such as FASTA or MEGA. Therefore, we use a simple XML format that allows the batch processing of hundreds of strains at one time (Figure 2).

Formatting the input data with a basic XML wrapper around a set of seven sequences for each strain allows a user to produce a file, for an unlimited number of strains, that can be used for batch processing. To aid production of such a format, each of the MLST species subsites at <http://www.mlst.net> provides a modified Access database that allows users to store their sequence and strain information in one place, and allows the data to be exported in bulk in the correct XML format, without the need for a user to manually produce the document. Furthermore, for sequencing laboratories using the STARS (<http://www.molbiol.ox.ac.uk/~paediat/stars/>) platform for MLST, we provide a facility to convert the FASTA files generated by STARS into the XML format via a web form.

When a user uploads the generated XML file (Figure 1D), the sequences for each of the seven loci in all of the strains are checked for invalid characters and correct length. Each sequence is then queried against the appropriate allele database (Figure 1E). If found, the allele number is returned and, if unknown, the user can look further into the sequence differences between the query allele and the most similar alleles in the database (Figure 1C). If all the seven loci are found, the allelic profile of the strain is queried against a look-up table of STs within the database and, if a match is found, the ST number is returned. If the allelic profile is previously unknown this information is returned. The batch procedure, therefore, automatically returns a table with the alleles, allelic profiles

Figure 1. Schematic representation of a typical MLST workflow. (A) Sequences can be entered locus by locus (single locus query), or all seven loci from a single strain (multiple locus query), or (D) by uploading a XML file with a set of strains and their sequences (batch strain query). (B) For the single and multiple locus queries, sequences that are not in the database are identified, and can be compared with the sequences of all known alleles using Jalview (7), or (C) the nucleotide differences compared with the most similar alleles can be displayed. The batch strain query (D) returns a strain table (E), which shows the allele number for each locus if known and the allelic profile if all the seven alleles and the ST are known. Strains that have the most similar allelic profiles to query strains are displayed as a table or by cluster analysis (F), and further information about them can be obtained. For the pneumococcal example used here, the query strains can be compared with the reference set of pneumococcal strains and closely related streptococcal strains, to establish whether or not they are pneumococci, using the concatenated sequences to construct a neighbor-joining tree (H). The relationship of unknown strains to the whole population can also be investigated using eBURST (G).

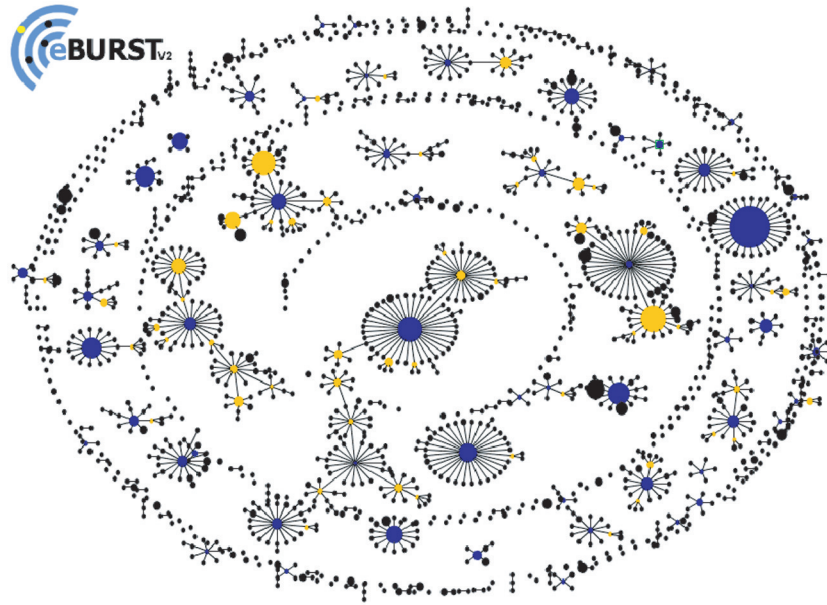


Figure 3. A population snapshot of the entire *S.pneumoniae* MLST database showing all major and minor clonal complexes viewed using eBURST.

and STs of all the input strains, flagging up those alleles and STs that are previously unknown (Figure 1E).

Comparing query strains to the database: clustering using allelic profiles

The simplest approach is to identify those strains in the database that have some minimum level of similarity in their allelic profile to each query strain (e.g. sharing alleles at ≥ 4 of the seven loci), and to show the relationship of the query strain to those returned from the database query using a dendrogram, based on the matrix of pairwise differences between the allelic profiles of the strains (Figure 1F).

Comparing query strains to the database: using eBURST

Traditionally, dendrograms have been the method of choice for displaying the implied relationships between strains of a bacterial population or species. However, although dendrograms are good at visualizing the clusters of identical or very similar strains, the bifurcating process of lineage splitting implied by a dendrogram is a very poor representation of the way in which bacterial lineages emerge and diversify. A new algorithm, BURST (10), was recently introduced that does not impose a tree-like pattern of descent, but rather uses an appropriate model of recent bacterial evolution. In addition, it is very difficult to display the relatedness of all strains in a large MLST database, including thousands of STs, on a dendrogram, and better ways of displaying the relationships among all strains in large MLST databases are required.

Briefly, the model incorporated into BURST assumes that, due to selection or genetic drift, some genotypes will occasionally increase in frequency in the population and will then gradually diversify by the accumulation of mutation(s) and/or recombinational replacements, resulting in slight variants of the founding genotype. Initially, members of this emerging clone will be indistinguishable in allelic profile by MLST,

however with time, the clone will diversify to produce a number of variants in which one of the seven MLST loci has been altered—single locus variants (SLVs). Further diversification will produce variants of the founder ST that differ at two out of the seven loci—double locus variants (DLVs). In this simple model, bacterial populations will consist of a series of clonal complexes (sets of variants of a founding genotype) that can be recognized from the allelic profiles of the strains within a MLST database (10).

An interactive implementation of the BURST algorithm, eBURSTv2 (10), is integrated within the MLST websites at <http://www.mlst.net> as a JAVA™ applet and can be used to explore the relationships among strains within the database and to explore the relationships of newly characterized strains to those in the database (Figure 1G). eBURST uses the STs and their associated allelic profiles as input and, using the default setting, divides the strains into groups in which all STs in the same group share ≥ 6 out of 7 loci with at least one other member of that group, resulting in non-overlapping groups or clonal complexes. Of particular value is the ability to link back to the MLST database from the eBURST diagram of a clonal complex, and the ability to display all the STs in a large MLST database in a single diagram [(10); a population snapshot; and Figure 3], showing all the major and minor clonal complexes, and individual STs that are relatively distantly related to all other STs.

Comparing query strains to the database: using the concatenated sequences

The ability to concatenate the sequences at the seven loci, maintaining the correct reading frame, and to construct a neighbor-joining tree based on these sequences is provided, but needs to be used with considerable caution. A module from MEGA (11) provides the tree topology in Newick format which is then displayed using the ATV applet (12). Allelic changes at the MLST loci will occur (to a varying degree

depending on the species) by recombination, and in many cases the relative contribution of recombination and point mutation to the diversification of strains will be unknown (13). A long history of recombination will preclude the recovery of the true phylogenetic relationships between distantly related bacterial strains and even the relatedness between similar strains may be better represented on a tree based on differences in allelic profiles than one based on differences in the concatenated sequences. However, there are specific issues that can be usefully addressed by using the concatenated sequences. For example, the *Burkholderia pseudomallei* database includes strains of closely related species and the *B.pseudomallei* MLST website provides a facility to examine the position of a query strain on the tree constructed using concatenated sequences, which can establish whether the query strain is *B.pseudomallei* or something similar to, but distinct from, *B.pseudomallei* (14). Similarly, there is considerable confusion about whether strains that appear to be *Streptococcus pneumoniae*, but which cannot be assigned to a pneumococcal capsular serotype, are authentic pneumococci that do not produce a capsule or are members of a similar but distinct streptococcal population. The pneumococcal MLST website has a facility to examine whether a query strain clusters within a reference set of *S.pneumoniae* strains, or with the related population, using a tree based on concatenated sequences, which can resolve this issue in most cases (see the following section; Figure 1H). Trees based on concatenated sequences may also be useful for assigning *Haemophilus influenzae* strains to major lineages (15) or for *Staphylococcus aureus* where recombination appears to be rare (16).

Typical workflow for data entry using the batch strain query

Here, we consider the workflow of a user analyzing a number of recently sequenced strains using batch entry. As an illustrative example we focus on a single representative mlst.net species website, <http://spneumoniae.mlst.net>, the site for characterizing strains of *S.pneumoniae* (17).

The uploaded XML file of a batch of *S.pneumoniae* strains and their associated sequences results in a table of results (Figure 1E). Error messages (red) alert the user to the fact that some sequences are of the wrong length for that locus (strain 8) or contain unexpected characters (strain 13). In some strains, all the alleles are previously known and the allele numbers are returned in the results table. For some of these strains, the combinations of alleles at the seven loci (allelic profiles) are also known and the ST number is shown in the table (e.g. strain 4). In one case (strain 14) the alleles are all known but the combination of alleles is previously unknown. In other strains, one or more alleles are unknown and the ST must also be unknown (e.g. strain 3), and the ST is flagged as incomplete, as the new alleles have to be checked and assigned new allele numbers by the curator. Clicking on 'unknown' allele highlights the nucleotide differences in the new allele compared with the most similar alleles (Figure 1C).

None of the alleles in strain 15 are found in the *S.pneumoniae* database, and there is therefore some uncertainty whether this strain is a pneumococcus. To investigate the status of this strain further, the user can select the option to examine the phylogenetic status of the strain, by using the concatenated

sequences to compare its position on a reference tree (Figure 1H), which includes a set of strains covering the known diversity of authentic pneumococci, and a set of closely related strains that are similar to but distinct from the authentic pneumococci (W. P. Hanage and B. G. Spratt, unpublished data). The sequences of the loci of the query strain are concatenated, and the sequence is added to a stored file containing the concatenated sequences of the reference strains, and a neighbor-joining tree is constructed (Figure 1H). Using this approach, strain 15, which has an unknown allelic profile but known alleles at all loci, clusters within the authentic pneumococci, but strain 14 with new alleles at all loci is clearly not a pneumococcus, as it clusters away from the pneumococci and within the more diverse set of related streptococcal strains.

From the results of the batch strain query, the user can also relate their unknown STs to all other strains in the MLST database using eBURST (Figure 1G). The unknown STs are assigned unique temporary ST numbers, to distinguish them from the STs in the database. In Figure 1G, strain 14 has been assigned the temporary ST10001 and by eBURST it can be seen to be a SLV of ST156 within one of the major pneumococcal clonal complexes. Any strain in the batch strain query (excepting those with alleles of incorrect length or with unexpected characters) can be compared with the MLST database as, using the eBURST option, new alleles, as well as new STs, are given temporary numbers allowing them to be analyzed by the program.

CONCLUSIONS

Websites for evaluating the taxonomic status of strains using 16S rRNA sequences are well established and, in recent years, several websites have been developed for molecular epidemiology and population genetics, to assign isolates of bacterial species to strains, lineages and clonal complexes, using data generated by MLST. We describe the set of MLST species websites within <http://www.mlst.net>, and the tools that allow users to identify query strains, and to explore their relationship with other strains in the database. MLST is being widely used and there is a need for new ways to input and query large sets of strains and to display the relatedness of the many thousands of strains within the larger MLST websites. Some progress has been made to achieve these aims and in future we envisage a fully automated procedure, with data flowing directly from sequencer to ST assignment. Those developing new MLST schemes for bacterial or fungal species can join <http://www.mlst.net> to take advantage of the features available at this site, and to have a consistency of format for the MLST websites. A slightly different common format for MLST websites is provided by those species sites (such as that for *Neisseria meningitidis*) hosted at <http://pubmlst.org>. Hosting of new MLST schemes at <http://www.mlst.net> allows the databases to be stored and backed up on servers at Imperial College London but with remote strain entry, ownership and curation, by the developer of the MLST scheme, using MLST curation software.

ACKNOWLEDGEMENTS

We acknowledge the contribution of Man-Suen Chan who developed the initial MLST website software, James Abbott

for the STARS XML conversion script, and Sudhir Kumar and Joel Dudley for providing the server-side MEGA module for concatenated tree drawing. The development of <http://www.mlst.net> was supported by the Wellcome Trust. Part of the publication cost for this article was provided by JISC. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA*, **95**, 3140–3145.
2. Hanage, W.P., Feil, E.J., Brueggemann, A.B. and Spratt, B.G. (2004) Multilocus sequence typing: strain characterization, population biology, and patterns of evolutionary descent. In Persing, D.H., Tenover, F.C., Versalovic, J., Tang, Y., Unger, E.R., Relman, D.A. and White, T.J. (eds), *Molecular Microbiology: Diagnostic Principles and Practice*. American Society Press, Washington DC, pp. 235–243.
3. Urwin, R. and Maiden, M.C. (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.*, **10**, 479–487.
4. Bournonville, M.E., Aanensen, D.M., Morand, S., Theraud, M., Spratt, B.G. and d'Enfert, C. (2004) Multilocus sequence typing of *Candida albicans*: data exchange and applications. *Infect. Genet. Evol.*, **6**, 243–252.
5. Dodgson, A.R., Pujol, C., Denning, D.W., Soll, D.R. and Fox, A.J. (2003) Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. *J. Clin. Microbiol.*, **12**, 5709–5717.
6. Chan, M.S., Maiden, M.C.J. and Spratt, B.G. (2001) Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics*, **17**, 1077–1083.
7. Jolley, K.A., Chan, M.S. and Maiden, M.C.J. (2004) mlstdbNet—distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics*, **5**, 86.
8. Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
9. Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **3**, 426–427.
10. Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P. and Spratt, B.G. (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.*, **186**, 1518–1530.
11. Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. *Bioinformatics*, **17**, 1244–1245.
12. Zmasek, C.M. and Eddy, S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
13. Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E. *et al.* (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl Acad. Sci. USA*, **98**, 182–187.
14. Godoy, D., Randle, G., Simpson, A.J., Aanensen, D.M., Pitt, T.L., Kinoshita, R. and Spratt, B.G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J. Clin. Microbiol.*, **5**, 2068–2079.
15. Meats, E., Feil, E.J., Stringer, S., Cody, A.J., Goldstein, R., Kroll, J.S., Popovic, T. and Spratt, B.G. (2003) Characterization of encapsulated and non-encapsulated *Haemophilus influenzae*, and determination of phylogenetic relationships, using multilocus sequence typing. *J. Clin. Microbiol.*, **41**, 1623–1636.
16. Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, A., Peacock, S., Maynard Smith, J., Murphy, M., Spratt, B.G. *et al.* (2003) How clonal is *Staphylococcus aureus*? *J. Bacteriol.*, **185**, 3307–3316.
17. Enright, M.C. and Spratt, B.G. (1998) A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology*, **144**, 349–360.